Preventive Veterinary Medicine 166 (2019) 39–48

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

journal homepage: www.elsevier.com/locate/prevetmed

Drivers for the development of an Animal Health Surveillance Ontology (AHSO)

Fernanda C. Dórea[a,*,1], Flavie Vial[b], Karl Hammar[c,d], Ann Lindberg[a], Patrick Lambrix[d,e], Eva Blomqvist[d], Crawford W. Revie[f]

[a] Department of Disease Control and Epidemiology, National Veterinary Institute, Sweden
[b] Epi-Connect, Skogås, Sweden
[c] Department of Computer Science and Informatics, Jönköping University, Sweden
[d] Department of Computer and Information Science, Linköping University, Sweden
[e] Swedish e-Science Centre, Linköping University, Sweden
[f] Atlantic Veterinary College, University of Prince Edward Island, Canada

ARTICLE INFO

Keywords:
Syndromic surveillance
Classification
Vocabulary
Terminology
Standards

ABSTRACT

Comprehensive reviews of syndromic surveillance in animal health have highlighted the hindrances to integration and interoperability among systems when data emerge from different sources. Discussions with syndromic surveillance experts in the fields of animal and public health, as well as computer scientists from the field of information management, have led to the conclusion that a major component of any solution will involve the adoption of ontologies. Here we describe the advantages of such an approach, and the steps taken to set up the Animal Health Surveillance Ontological (AHSO) framework. The AHSO framework is modelled in OWL, the W3C standard Semantic Web language for representing rich and complex knowledge. We illustrate how the framework can incorporate knowledge directly from domain experts or from data-driven sources, as well as by integrating existing mature ontological components from related disciplines. The development and extent of AHSO will be community driven and the final products in the framework will be open-access.

1. Introduction

In 2011, Dórea and collaborators (Dórea et al., 2011) provided a comprehensive review of syndromic surveillance in animal health, highlighting ongoing initiatives and opportunities for automated extraction of surveillance information from the rapidly growing quantity of computerized animal health data. An update of that review in 2016 (Dórea and Vial, 2016) indicated remarkable growth in the field, but concluded that automated analysis and interpretation of animal health data was still hindered by a number of limitations. In particular, the lack of syndromic classification standards was preventing integration and interoperability among systems using different data sources.

The issue of compatibility becomes increasingly relevant as the number and type of animal health data sources grows, as do the opportunities and pressure for surveillance officials to gather (timely) evidence from multivariate surveillance systems (Gates et al., 2015; VanderWaal et al., 2016). Secondary use and interpretation of data collected in different contexts is only possible if "the integrity and meaning of the data is preserved throughout the integration process", a quoted definition of semantic interoperability (Al Manir et al., 2018).

The adoption of data coding standards, such as the Systematized Nomenclature of Medicine (SNOMED) (Stearns et al., 2001) can improve interoperability. However, as pointed out by Dixon et al. (2014), this puts the burden of data reuse solely on the data providers. The authors suggest, instead, a cooperative approach and maximization of the value of data collected through modern information management systems. Mirhaji (2009) listed interoperability and multidisciplinary reuse as two of four enabling principles to achieve translational public health informatics. The other two, dynamic adaptability and human-computer interaction, will require dynamic knowledge models that can be used by humans and computers to reason with large volumes and variety of data.

Considering these points, in this paper we outline the development of an ontological framework to promote semantic interoperability among health sources to be used for syndromic surveillance. Rather than coding data, we suggest gathering knowledge from the community

* Corresponding author at: Department of Disease Control and Epidemiology. National Veterinary Institute (SVA), Uppsala, SE-751 89, Sweden.
E-mail address: fernanda.dorea@sva.se (F.C. Dórea).
[1] Current affiliation: National Wildlife Management Centre, Animal and Plant Health Agency, York, UK.

https://doi.org/10.1016/j.prevetmed.2019.03.002
Received 5 March 2018; Received in revised form 7 January 2019; Accepted 5 March 2019

of domain experts to develop harmonized rules to interpret data, that is, to translate health data into syndromic representations.

2. Ontologies – what and why?

Ontologies are data models which capture, in a way that is transparent to both humans and digital devices/agents, the knowledge structures needed to address tasks in a specific context. They therefore facilitate communication among humans, and provide interoperability among systems (machines) (Lambrix and Strömbäck, 2007).

Consider any specific knowledge set – for instance the biomedical knowledge involved in analyzing health records for the purpose of syndromic surveillance. Experts can agree that a case of abortion in pigs, or the record of a suspicion of brucellosis, can both be classified as "reproductive syndrome" events. This simple example involves a number of concepts: the organism "*Brucella suis*", the disease "brucellosis", the clinical sign "abortion", among others. It also involves a number of relationships (underlined), such as: "*Brucella suis*" is a "bacterium", which causes the disease "brucellosis", which affects the "reproductive system", and can have clinical sign "abortion".

Traditional vocabulary agreements and terminologies provide lists of predefined concepts. Examples would be a hierarchical list of organisms and their taxonomic classification (which would contain *Brucella suis* as a species of the genus Brucella), or a list of anatomical entities. Ontologies are machine-interpretable models that include the semantic relationships among concepts (Noy and McGuinness, 2001). That is, they can capture both the terminologies and the relationships across them, storing them in a format that can be used by machines to reason with the data. Rather than expect data to be "smart", these models enable smart applications, which can get the right data to the right place, and extract information from them (Allemang and Hendler, 2011). We can for instance construct syndromic definitions in real-time, rather than requiring that data be coded according to specific syndromic standards, and change these definitions as knowledge evolves or new threats emerge. Whether we define a respiratory syndrome based on clinical signs, anatomical location of pathological condition, or both, we can query data which has not been specifically labelled with any particular syndromic definition. This facilitates the automation of many tasks, enhances interoperability among systems (including historical/legacy systems), and increases the amount of information that can be extracted from raw data.

Previous work has pointed out the limitations of biomedical terminologies which only contain hierarchical relationships, or contain relationships that are too generic (Ceusters et al., 2003). MeSH terms (Medical Subject Headings; Lipscomb, 2000), for instance, make no distinction between the relationships is a and part of (Ferreira et al., 2012) (an example of this difference in a medical context is: lung is a organ, which is part of the lower respiratory tract). In SNOMED-CT, codes are organized into a hierarchy, but the relationships between subclasses and the class they specialize does not always hold to the formalism of an "is-a" relationship. Under clinical findings, for instance, there is a code for "doctor left practice". Under "body structure", there is a code for "normal anatomy" (i.e. reasoning with this knowledge would require us to state that "normal anatomy" is a "body structure"). Data annotated following terminologies lacking proper semantic logic can be queried using the concepts in the terminology, but reasoning with the data to discover new relationships and draw inferences will be limited by any such inconsistencies in their semantic structure (Pesquita et al., 2014).

Transforming data into actionable information is the goal of any data analysis carried out in surveillance. A machine interpretable model of the knowledge needed to interpret health data for surveillance is an important step towards allowing computers to process large volume and variety of data. Humans can then focus on digesting the processed information and taking decisions, while the timeliness of the overall process is improved. The benefits of an approach based on a computer-interpretable knowledge model include:

- It provides a transparent and common understanding of the concepts documented in the ontology, including but not restricted to syndromes.
- Data sources do not need to be coded according to specific standards. Institutions may continue to use their own individual coding practices. Data can then be marked up (often in a semi-automatic manner) to allow for querying via the ontology.
- Since the data does not need to be coded into specific syndromes, the parameters of the search are defined based on current needs. Today it may be syndromes, but tomorrow it could be a search focused on specific clinical signs known to be associated with an emerging disease.
- It can easily accommodate knowledge change or new knowledge, which tend to be especially important in the case of emerging disease detection.
- It allows information to be queried based on relationships between concepts, for instance "all diseases which can cause clinical sign A", or "all clinical signs associated with disease X".
- Knowledge reuse – not only among syndromic surveillance initiatives, but especially by incorporating existing knowledge already contained in other ontologies.

3. Animal Health Surveillance Ontological framework (AHSO)

3.1. Animal health surveillance context

Ferreira and collaborators have pointed out the complexity and multidisciplinary nature of epidemiology, highlighting the need for an integrative framework that can only be addressed by enabling semantic technologies (Ferreira et al., 2013 2012; Pesquita et al., 2014). Surveillance in general, and syndromic surveillance in particular, face the additional challenge of secondary use of data. Information for decision making must be extracted from data that were collected for alternative purposes, including clinical records, laboratory findings or slaughter inspection data.

Digitalized data about an animal or herd can be collected along the entire cycle of animal production, even in the absence of disease events, as summarized in Fig. 1-B. Health care encounters can generate additional records, such as clinical and laboratory data. As schematized in Fig. 1-A, health events are not directly modelled in any of the sources of data available. Surveillance makes use of recorded observations which can be, often, related to the same underlying health event. Health events can be a disease occurrence or a regular animal production cycle observation, such as birth or product yield (e.g. milk, weight gain, etc). To be able to interpret these observations, knowledge models capable of automated information extraction need to account for the structure of animal production, the nature of the observation context and data being recorded, in addition to the relevant health information (Fig. 1-C).

Modelling the knowledge needed to interpret information from each of these data sources (the various observation contexts illustrated in Fig. 1-C) requires that concepts from a multitude of disciplines and specific subject fields be addressed, in addition to establishing relationships among these concepts and contexts.

Ontologies provide the ideal framework for building knowledge models that are interoperable and reusable, as demonstrated by the growth and success of the Open Biomedical Ontologies (OBO) Foundry initiative (Smith et al., 2007). We are addressing this task by reviewing and reusing available knowledge models, and building a community of experts to address knowledge gaps and create novel conceptual mappings.

3.2. Modular development: from data to a data model

AHSO has been designed as an ontological framework, rather than a

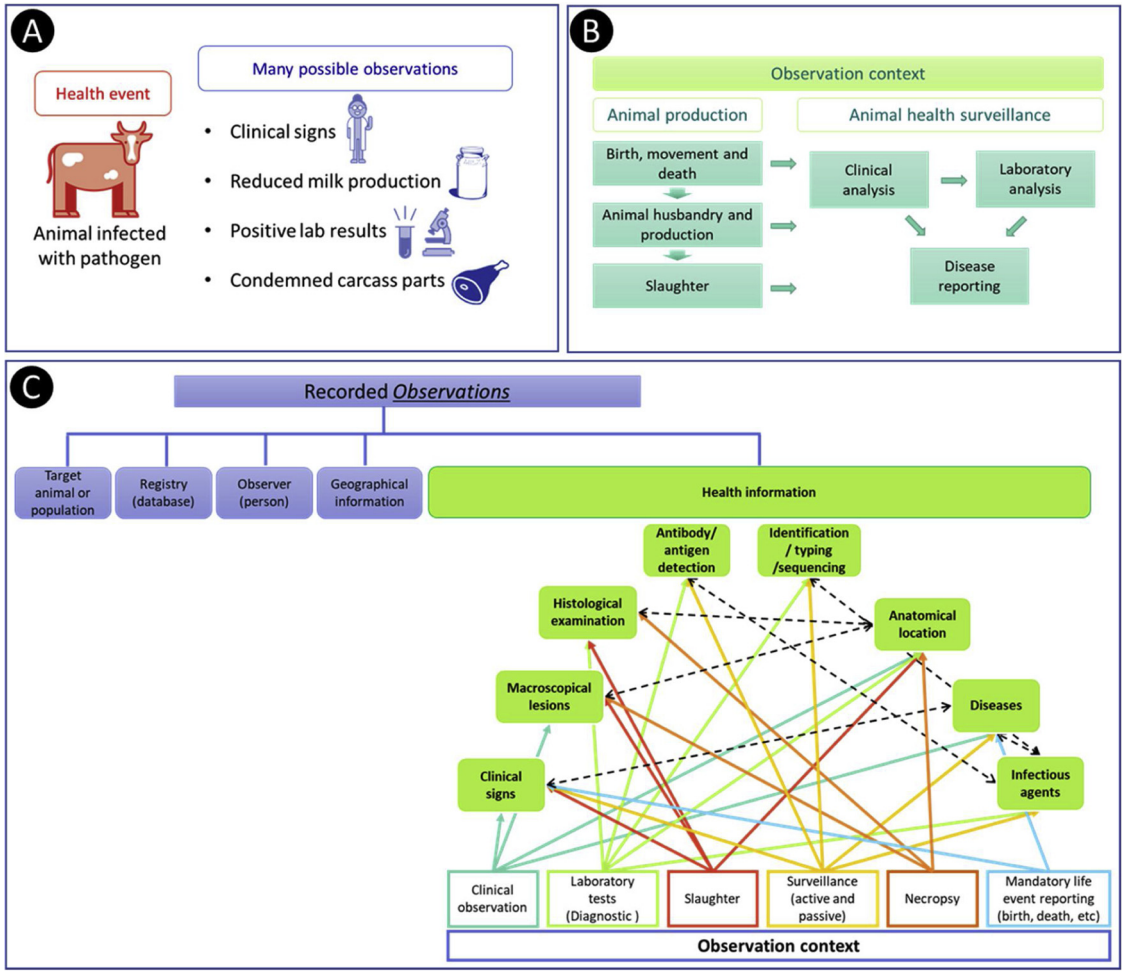


Fig. 1. Animal health surveillance information context. A) The AHSO framework makes the specific assumption that the data to be processed are composed of (potentially multiple) recordings of observations made about an underlying health event, which is not directly modelled. B) These observations can be recorded in different contexts, as part of the routine animal production cycle, or specifically triggered by health events. C) The AHSO framework has a core that tries to capture the structure of the animal production (blue boxes). Modeling the health information will require a number of knowledge models, for different observation contexts, which the framework will connect.

single ontology. A key component of AHSO will be the provision of a flexible structure that can be utilised to connect the various pieces of existing knowledge required to produce actionable information for surveillance.

In its core, the framework provides a structure to model the contextual information that comes with every observation of a health event: information about the target population, such as population unit (herd or animal, for instance), animal species, breed, age and sex; geographical information; information about the observer; about the registration or context in which an observation was made. The latter is important for instance to record whether a data observation was triggered by a health event (such as a visit to the veterinarian), or was part of a routine data recording event, as would typically be the case for production data. It also supports tasks such as: identifying mandatory data recordings (e.g. cattle movement in Europe), inferring the technical level of the observer (e.g. veterinarian versus data owner), and determining the specificity level of the health information (e.g. laboratory results versus clinical finding or presenting complaint).

The initial core framework has been modelled in OWL (Web Ontology Language), using Protégé 4.1 (https://protege.stanford.edu/) (Musen, 2005).

Given the magnitude and complexity of the task, we have chosen to develop the ontology using a data-driven approach. One prototypic data source is addressed at a time, and new concepts are added to the ontological framework using a combination of top-down structuring, and bottom-up generation of concepts from data. For example, modelling diagnostic codes used by pathologists performing necropsies requires, at a minimum, that the concept of a pathological lesion (such as, for instance, inflammation) be modelled and that it be possible to define an anatomical location (say, the lungs). As we further inspect data, we may find the need to also model concepts that allow for the identification of pathogens responsible for specific pathological conditions. In the OWL language, concepts are modelled as “classes” which can have “sub-classes”. The classes required are added to the framework – see for instance the green boxes in Fig. 1-C. The next step is then to search for existing ontologies that may already contain the classes, or have at least defined vocabularies and terminologies. In the absence of such, the novel classes are generated from data source examples, using cycles of automated learning from data and expert review. Methods are illustrated for an example scenario below.

Addressing each specific data source results in functional modules which can already be used to automate surveillance information extraction in specific contexts. As more modules are added, the ontology will become useful in a greater number of contexts.

The addition of new modules builds on existing knowledge already available within the ontology, and many of the biological concepts can be reused across several modules; for instance, the concept of “anatomical location”.

3.3. Reusing knowledge

Ferreira et al. (2012) conducted a review of sources for modelling epidemiological knowledge, and concluded that no existing ontological framework provided the necessary characteristics to enable annotation of epidemiological data. However, the authors identified a number of sources of knowledge that contain important concepts that could be reused, such as SNOMED (Stearns et al., 2001), the Unified Medical Language System, UMLS (Bodenreider, 2004) and Medical Subject Headings, MeSH (Lipscomb, 2000). The authors also noted a number of biomedical ontologies which can model specific concepts within the epidemiological framework, such as ontologies covering diseases, clinical signs, pathogens as well as a number of anatomical ontologies. Their results support our approach of establishing a framework that is designed to connect existing resources.

We reviewed all biomedical ontologies listed in the OBO Foundry (http://www.obofoundry.org/) and the BioPortal repository of biomedical ontologies (http://bioportal.bioontology.org/) to identify ontologies that could potentially contribute to the AHSO framework. In order to identify terminology resources which are already in use by the surveillance and scientific community, and could be adapted into an ontological framework, we continuously asked for feedback from these communities, at relevant conferences and within our project networks.

We identified 42 resources related to the health surveillance domain (see Supplementary Material 1). From those, 24 were considered relevant to contribute, in some manner, to the AHSO framework, and they are listed in Table 1. Four are ontologies which contain useful concepts, but do not appear to be under active development, and therefore we would need to import their concepts and review the associated structure. Another 11 resources contain useful concepts, but are terminologies or vocabularies not constructed using a semantic framework, or are not available in OWL, and therefore cannot be readily imported into the AHSO framework. Nine ontologies available in OWL could contribute directly to the framework, and are highlighted in green on Table 1. Public tools exist which allow specific pieces (or even entire) ontologies to be imported for reuse. We highlight the web-based tool Ontofox (http://ontofox.hegroup.org) (Xiang et al., 2010). Reuse of existing ontologies is exemplified further below using the case of Uberon (a comprehensive animal anatomy ontology) (Mungall et al., 2012).

It is also worth highlighting GenEpiO (Genomic Epidemiology Ontology) (Griffiths et al., 2017), a recent initiative also focused on surveillance. As in the case of AHSO, GenEpiO is intended to provide a framework to integrate existing ontologies, and to develop new modules where needed. GenEpiO is developed primarily from the human surveillance perspective, focused on laboratory analysis (in particular next generation sequencing) and case reporting. In contrast, AHSO was motivated from the context of syndromic surveillance, and it is therefore primarily focused on secondary use of data. These two frameworks have the potential to be highly complementary, and as such AHSO will initially focus on modules not covered by GenEpiO, to avoid duplication of effort, and will explore opportunities for collaboration to allow combination of modules to/from GenEpiO at a later stage.

3.4. Building modules from data and expert elicitation

Not all required knowledge will be available in existing ontologies, nor will all relationships important to the process of animal health surveillance be adequately captured. When classes cannot be readily imported from other ontologies, they are created reviewing existing terminologies and coding the relevant concepts in OWL, within the AHSO framework. Ontology learning could also be employed in the future, by which new classes are created from available data through semi-automated processes (Buitelaar et al., 2005; Cimino, 1999; Donald et al., 2018). In either case, concepts added to the ontology are subjected to community review, as detailed in the section below.

We anticipate field experts to be an important source of knowledge for the ontology; however, no ontological design knowledge will be required of those making such contributions. Tools exist to collect information from experts in simple formats, such as Excel spreadsheets, from which the knowledge can be integrated into the ontology (Courtot, 2014). Project resources for community engagement are detailed further below.

4. Development of the framework core and content growth

To grow the ontology in modules, and also to create a workflow that does not rely on data sharing, we follow the guidelines of the eXtreme Design method (XD) (Blomqvist et al., 2016). Each iteration within the XD method is triggered by small specific examples of data which need to be modelled, and the process is focused on a test-driven and collaborative approach. In particular, the method is based on the use of ontology design patterns (ODP), which are “reusable modelling solutions that encode modelling best practices (Presutti et al., 2012)”. In other words, each cycle of development aims to solve a very specific modelling problem, and a catalogue of design patterns is searched to look for non-domain-specific modelling solutions which may be reused or adapted. An extended version of the XD method (Dragisic et al., 2015) also supports the integration of pieces of ontologies as well as the completion and debugging of the ontologies before and after integration with tools such as OOPS! (Poveda-Villalón et al., 2014) and RepOSE (Lambrix and Ivanova, 2013).

Based on a data/information inventory from interested partners, we have gathered documentation of 27 animal health data sources from 8 countries. Data-sharing is not always expected to be possible, but this should not hinder ontological development. The XD protocol accounts for development based on “requirement stories”, which are narrative examples of the events contained in a specific dataset, and therefore examples of the events that need to be modelled. Data owners can provide either a data sample or a user requirement story, and would therefore not be required to divulge potentially private data, nor would they require any understanding of the ontology building methodology.

Below is an example of a requirement story. It is meant to be a narrative version of a type of event that would be registered in a cattle movement registry (still-birth), in substitution to sharing data from such a registry:

> *Farmer Nilsson, during his morning visit of his stables in Skåne on the 10th of June 2015, notices that his cow Daisy gave birth during the night to a calf that was dead-at-birth. Farmer Nilsson notifies the abortion to the electronic cattle register.*

The requirement story informs questions around the types of concepts that will typically be needed to model the observation. In this example we would need to model the structure of animal production, such as animal clustering in herds, herd location, and ownership; as well as the fact that health information can be recorded at the animal or herd level. This story was chosen to drive structuring of the framework core, because it is focused not on the specific health information, but on the structure of animal production, animal ownership, and event reporting.

In traditional data recording, we often think of data as a two-dimensional spreadsheet. The concepts in the story above, such as the animal, the date and the occurrence, would be reported in specific columns. While simple, this format of data recording limits data reuse and creates common problems for data entry when a simple 2D format is inadequate to model the information requirements.

In the representation languages of the Semantic Web, such as OWL, each information point is considered individually as a triple of subject, predicate and object (Allemang and Hendler, 2011), for example: A (subject) P (predicate) B (object); where A and B can be concepts (e.g., cow, representing all cows) or instances (real world objects, e.g., Daisy, representing a particular cow). For instance, we can represent “cow is_a animal” and “Daisy has_birthdate 2011-03-23″.

Table 1
Inventory of ontologies and terminology resources that could be useful in the construction of the AHSO framework. White rows are ontological resources with high potential for direct reuse: the concepts are valuable for AHSO, available publicly and in the OWL language, and the ontology seems to be active. Rows shadowed gray are resources that may offer concepts for reuse, but they would need to either be coded into an ontological framework within the AHSO project (terminologies and vocabularies not available as an ontology) or updated within the project (inactive ontologies).

Ontology	OWL	Notes	Source and references
ARO	YES	ARO is a key component of the Comprehensive Antibiotic Resistance Database (CARD)	Part of the GenEpiO framework (https://genepio.org)
ATOL (Animal Trait Ontology for Livestock)	YES	Domains: Welfare, Growth and meat production trait, Mammary gland and milk production trait, eggs, nutrition, reproduction and fertility. Last updated in 2014.	https://bioportal.bioontology.org/ontologies/ATOL
BioCaster	YES	Designed specifically to mine news text online for event detection, so it only addresses some of the complexity needed.	https://github.com/nhcollier/biocaster-ontology
CMO (Clinical Measure Ontology)	NO	CMO is designed to standardize morphological and physiological measurement records generated from clinical and model organism research and health programs.	https://bioportal.bioontology.org/ontologies/CMO
EFSA's terminology catalogues	NO	The European Food Safety Authority provides several catalogues harmonizing terminology for collecting and analysis of data related to animal and food safety.	https://zenodo.org/record/344473#.XAi7mmhKiI5
EPO (Epidemiology Ontology)	YES	EPO is an ontology designed to support the semantic annotation of epidemiology resources, but it doesn't seem to be active or available online any longer.	http://www.obofoundry.org/ontology/epo.html
ESSO (Extended Syndromic Surveillance Ontology)	YES	An ontology for syndromic classification of emergency records and radiological diagnostics. Focuses on humans. It has not been updated since 2009.	https://bioportal.bioontology.org/ontologies/SSO
FoodOn (Food Ontology)	YES	A broadly scoped ontology representing entities which bear a "food role". Useful for modeling concepts in foodborne disease surveillance.	https://bioportal.bioontology.org/ontologies/FOODON
GenEpiO (Genomic Epidemiology Ontology)	YES	GenEpiO covers vocabulary necessary to identify, document and research foodborne pathogens and associated outbreaks. The ontology goals and project development structure are very aligned with AHSO, though the focus here is on humans and surveillance reports, rather than secondary use of data. We will try to focus on areas not covered by GenEpiO, so that we can build complementary, rather than overlapping modules.	https://bioportal.bioontology.org/ontologies/GENEPIO
ICAR (International Committee for Animal Recording)	NO	The International Committee for Animal Recording publishes standards for disease recording in livestock animals, as well as for the recording of animal genetic traits and production performance.	http://www.icar.org/

(*continued on next page*)

Table 1 (*continued*)

Ontology	OWL	Notes	Source and references
IDO (Infectious Disease Ontology)	YES	Suite of interoperable ontology modules that together cover the entire infectious disease domain.	https://bioportal.bioontology.org/ontologies/IDO
LOINC (Logical Observations, Identifiers, Names and Codes)	NO	LOINC is an international standard for identifying health measurements, observations, and documents. It is not in ontological format, but due to broad use and public availability, it is important to consider the concepts covered, and how they map to other resources.	http://loinc.org/downloads
MeSH (Medical Subject Headings)	NO	MeSH is the National Library of Medicine's controlled vocabulary thesaurus. It is a good source of concepts, but only hierarchical relationships are modelled. Some concepts are placed in multiple parts of the hierarchy.	http://www.nlm.nih.gov/mesh/MBrowser.html
MP (Mammalian phenotype ontology)	YES	Models abnormal morphology and physiology of specific systems.	https://bioportal.bioontology.org/ontologies/MP
NCBI (National Center for Biotechnology Information (NCBI) Organismal Classification)	NO	The NCBI Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This should be a reference for the listing of organisms, such as pathogens.	http://bioportal.bioontology.org/ontologies/NCBITAXON
OBI (Ontology for Biomedical Investigations)	YES	OBI is a model to report scientific investigations. The structure used to define assays, devices, objectives, etc, should be studied and if possible reused when modeling a surveillance system.	http://bioportal.bioontology.org/ontologies/OBI
PATO (Phenotypic Quality Ontology)	YES	Phenotypic qualities (properties). This ontology can be used in conjunction with other ontologies such as GO or anatomical ontologies to refer to phenotypes.	https://bioportal.bioontology.org/ontologies/PATO
PDO (Pathogenic Disease Ontology)	YES	An ontology to describe infectious diseases. It is focused on humans only, but should be compared to IDO and concepts reused when relevant.	https://bioportal.bioontology.org/ontologies/PDO
REPO (Reproductive Trait and Phenotype Ontology)	YES	Ontology for livestock reproductive traits and phenotypes. Not updated since 2011.	https://bioportal.bioontology.org/ontologies/REPO
RxNorm	NO	Standardized nomenclature for clinical drugs.	http://bioportal.bioontology.org/ontologies/RXNORM
SNOMED-CT	NO	A very extensive clinical health terminology. Relationships between concepts are only hierarchical.	http://bioportal.bioontology.org/ontologies/SNOMEDCT
SurvO (Surveillance Ontology)	?	SurvO will define key indicators of reportable disease surveillance systems. Still in the planning phase.	Planned as part of the GenEpiO framework (http://genepio.org)
UBERON	YES	Uberon is an integrated cross-species anatomy ontology representing a variety of entities classified according to traditional anatomical criteria such as structure, function and developmental lineage.	https://bioportal.bioontology.org/ontologies/UBERON
Ontology	**OWL**	**Notes**	**Source and references**
UMLS (Unified Medical Language System)	NO	This comprehensive collection of standards is broadly adopted and should be consulted as a source of terms for AHSO, and mapping to other resources also considered.	https://www.nlm.nih.gov/research/umls/

Our job when structuring the ontology is to model concepts in our knowledge domain as classes, where instances within these classes represent individual objects in the world that we wish to represent. For instance, Daisy, in the example above, may be an instance of the class "cow"; which could itself be a sub-class of the class "animal". We can then define the properties (which we have been calling the

relationships between concepts) that connect different classes in the world, by specifying their specific domain, range, and cardinality. For instance, we may wish to state that the "person" class can be linked to the "animal" class by the property "is owner of". We specify that the property has domain "person", and range "animal". We could actually set the range to any "animal population unit", and state that both "animal" and "herd" are ("is-a") "animal population unit". We then can set the cardinality – whether an animal owner must have at least one animal, and whether it can have more than one. Properties that link instances of two classes, such as this one, are called object properties. Data properties link an object to a value, for instance the observation "occurred on date".

The small requirement story above was subjected to several rounds of discussion among the authors, to model the structure of animal production in a way that would be robust enough to accommodate many different data sources. We needed to account for different formats of data recording, and various limitations that can be encountered, such as missing information regarding animals. It was agreed that it may often be difficult to identify the individual population units associated with each health event observation, and that much of the information gathered will be uniquely identified only at the level of the observation (unique observations in a dataset). This prompted the conclusion that all pieces of information must be linked to the observation ID, which resulted in the structure shown in Fig. 2-A. An example of how this model would inform the conversion of data into triples, using the requirement story, is also shown.

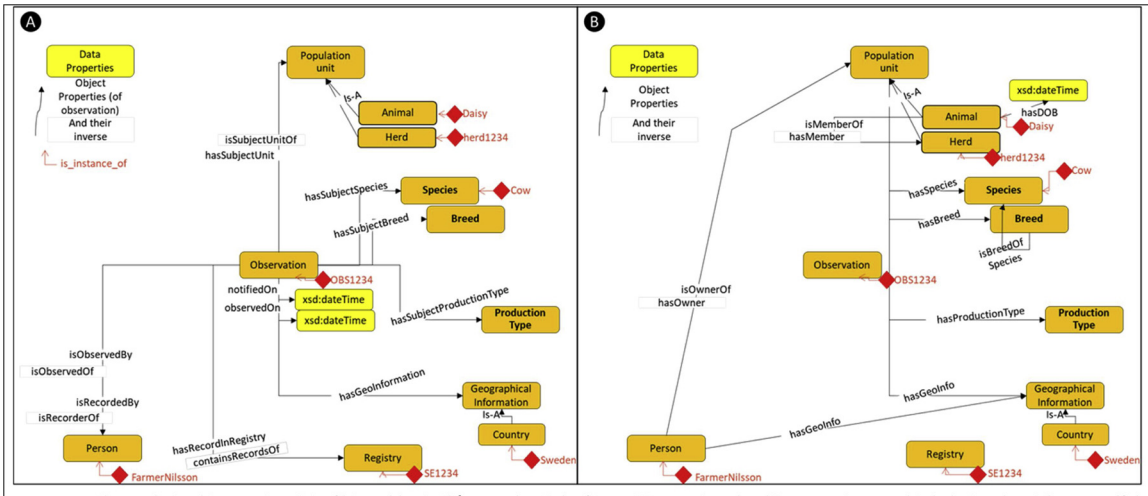
Additional properties linking the other modelled classes are shown in Fig. 2-B. Together, these figures show the entire structure of the current framework, that might be required to fulfil the relatively simple requirement story, though this structure is still missing the actual health information recording for the event. The model has grown to include concepts considered essential to model the basic animal production structure, which were not exemplified in the original requirement story, such as animal breed and the production type of the herd, as shown in Fig. 2.

In Fig. 2-A and -B, the non-specified relationships drawn as a hierarchical scheme (animal and herd to population unit) must be interpreted as an "is a" type relationship (an animal *is a* population unit). Each of the round-cornered boxes in Fig. 2 represents a top class that now needs to be specialized with subclasses representing the various concepts that need to be modelled, with varied depth of the hierarchy within each one. For instance the class "species" can be specialized using a complex taxonomy of animal species. For production type, we could find no ontology that addressed this specific concept, but many different standards relating to animal production can be found. We have thus chosen to model the concept using a data-driven approach, where new production types will be added as needed, to cover the modelling needs of each new data source (or requirement story) being addressed within each iteration of the AHSO development process.

By this stage of development, we have a framework of animal production, but as yet no model of health information. To test the process of specializing the classes learning from data, we explored a dataset of mandatory cattle movement reporting in Sweden. For this new iteration, we detailed the class "observation" with the following sub-classes: birth, death, declared movement, slaughter, and stillbirth. These are types of observations. Note for instance that stillbirth here serves as an observation trigger, rather than a clinical sign recorded by a veterinarian. This explicitly models the fact that the observation was a mandatory observation reported by a farmer; that is, the stillbirth was the trigger for the event. Specific additional classes will be needed later to model the observation of stillbirth as a clinical sign, by a veterinarian.

As a first step to tackling data containing diagnostic information, we considered the list of codes used in the pathology department of the Swedish National Veterinary Institute when annotating necropsy diagnostics (SVA-pathology dataset). This was chosen because the codes were relatively straightforward, often being in the form of "anatomical location" + "lesion type"; for example, "liver abscess".

Explicitly modelling such a set of codes around these two specific concepts facilitates data reuse of all the knowledge that has been coded around these concepts. Consider the specific example of pathology codes that are related to the respiratory system. In the SVA-pathology



Farmer Nilsson, during his morning visit of his stables in Skåne on the 10th of June 2015, notices that his cow Daisy gave birth during the night to a calf that was dead-at-birth. Farmer Nilsson notifies the abortion to the electronic cattle register.

This observation receives an unique identifier: OBS1234

OBS1234 is_instance_of Observation	Farmer_Nilsson isOwnerOf Daisy
OBS1234 observedOn 2015-06-10	Daisy is_instance_of Animal
OBS1234 hasSubjectUnit Daisy	Farmer_Nilsson isOwnerOf herd1234
OBS1234 has SubjectSpecies Cow	Daisy isMemberOf herd1234
OBS1234 hasCountry Sweden	Daisy hasSpecies Cow
OBS1234 hasRecordinRegistry SE1234	
OBS1234 isObservedby Farmer_Nilsson	

Fig. 2. AHSO core classes and relationships. Concepts are shown in round-cornered boxes, and relationships as arrows. Panel A shows relationships directly linked to the class "Observation" (and sub-classes); panel B shows other relationships. The requirement story exemplified in the text is displayed, which instances in that story shown in red. Relationships from both panels are exemplified using those example instances.

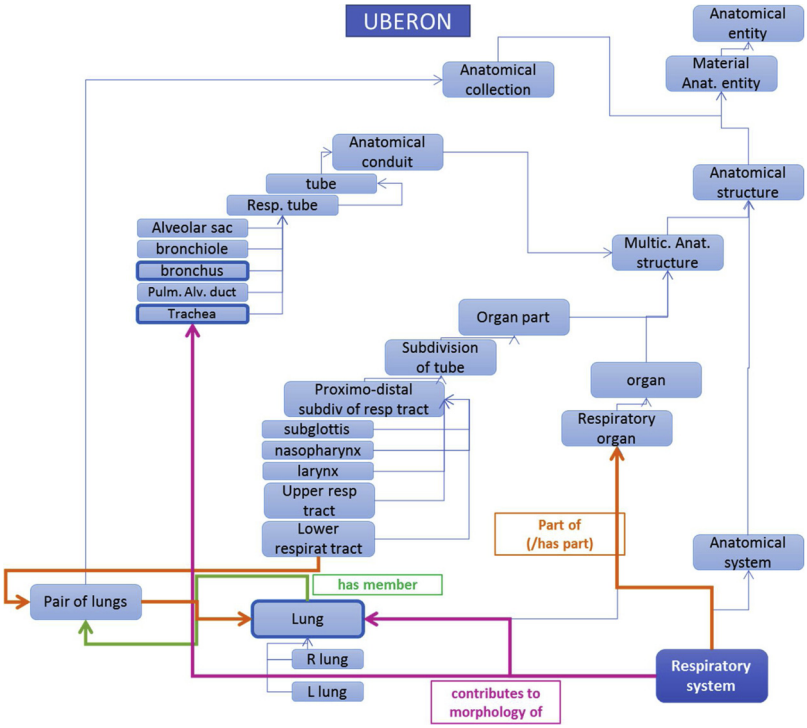


Fig. 3. UBERON (anatomy ontology) classes linked to the anatomical entities "lungs", "bronchus" and "trachea". Blue arrows are to be understood as IS-A relationships.

dataset we found the following anatomical locations mentioned among the various diagnostic codes: pleura and thorax; sinuses and air sacs; trachea and bronchus; nose; and lungs. Fig. 3 shows a scheme of the many concepts linked to the anatomical entities "lungs", "bronchus" and "trachea" in the Uberon ontology (Mungall et al., 2012). If we map all diagnostic codes involving lungs to the specific "lungs" class in Uberon (for instance explicitly saying that a code for pneumonia "has_anatomical_location" lungs), a query would be able to return observations of "pneumonia" automatically when a user requests any lesions in the lower respiratory tract.

Consider this list of specific diagnostic codes relating to the respiratory system in the SVA-pathology dataset:

1) Catarrhal pneumonia
2) Purulent pneumonia
3) Mycotic pneumonia
4) Respiratory syncytial cell pneumonia (RSV)
5) Acute lung emphysema
6) Normal lungs

As we inspect these codes in further detail, we realise that modelling the concepts of a lesion, or a "pathological process" and an "anatomical entity" may not be enough. "Normal lungs", for instance, have no pathological process. Yet, we need to model the concept of a normal lung, as observing no lesion is different from not making an observation in the lungs. One solution is to consider pathological codes as belonging to a broader category of "health findings", and model the facts that health findings can have an observed "morphological description" (for instance normal morphology), a "pathological process" (for instance inflammation), and a specific "anatomical location".

The class "anatomical entity" was created to represent the latter, and imported from the Uberon ontology. Uberon is an extensive ontology that also includes anatomical details of invertebrates, reptiles and amphibians. To take advantage of the relevant complexity in Uberon, without importing a set larger than necessary, we created a subset of Uberon using the following steps: classes in Uberon representing each anatomical location found in the SVA-pathology dataset were listed; OntoFox was then used to generate a subset of Uberon that would contain the classes in this list, using the option "includeAllAxioms". This option ensures that classes are imported with all of their annotations (definition, comments, references to other ontologies, synonyms, translation to other languages, etc), together with all properties involving these classes, and the classes linked to them through such properties. This results, for instance, in importing all of the classes and properties depicted in Fig. 3, when OntoFox was instructed to import only the three classes highlighted in thicker outline ("lung", "bronchus" and "trachea").

Looking at terms 3 and 4 in the list above, we realise that it is now necessary to create classes for organisms and the specific diseases caused by them. Term 5, "acute lung emphysema", highlights the need to model various qualifiers for the pathological conditions, such as acute and chronic. Modelling these concepts and classes are part of the on-going ASHO design process.

In addition to integrating general and widely-used ontologies, such as Uberon, within the AHSO framework we also wish to explore the use of controlled vocabularies that specifically target the veterinary context. Our initial focus will be the Veterinary Nomenclature (VeNom) clinical terminology (http://www.venomcoding.org). These codes are being considered in parallel to the pathology codes, to ensure that we can model pathological conditions without confusing these with the concept of observational context. That is, some of the findings from necropsy cases could also be findings reported by clinicians, based on simple clinical evaluation or further tests. Although a veterinarian can give a diagnosis of "pneumonia", this should not be listed in the ontology as a "clinical finding" concept – pneumonia is the pathology, and the observation context is clinical finding. Whether a health finding has been reported by an animal owner (presenting complaint), a physician (clinical finding) or a pathologist during necropsy, is a matter of the observation context. Codes for pneumonia should not be repeated, for instance, under different classes for each of these contexts.

Examples such as this prompted us to separate what was being modelled simply as "observation", into "observation" and "observation

context". Under the "observation context" of a veterinary visit, for instance, there could be observations of "clinical history", "presenting complaint", "physical examination", "clinical finding", among others. This provides an example of how complexity may be gradually added to the ontology as we address a more complete set of data examples.

5. Community involvement

We are developing AHSO aiming for compliance with the OBO Foundry's (evolving) set of principles, which include open use, collaborative development, non-overlapping and strictly-scoped content, and common syntax and relations. AHSO is available publicly in the BioPortal (http://bioportal.bioontology.org/ontologies/AHSO).

To ensure open development, we have created four main Web resources:

- Besides being accessible in the BioPortal, the latest ontology release can also be found at http://datadrivensurveillance.org/ahso. The page is managed with content negotiation, so that people visiting this address will find a webpage with information regarding the project. However, the URL can also be accessed directly using ontology programming software (such as Protégé) to interact directly with the ".owl" ontology file.
- GitHub (https://github.com/SVA-SE/AHSO) is being used to publicly store ontology source files, which allows any interested parties to: see and download the current version of the ontology; suggest improvements and corrections; submit requirement stories or other issues that need to be addressed in future development stages; access a wiki with relevant references regarding ontologies in general, and the project in particular.
- A discussion forum (Google group) has been created to establish conversations with a community of users particularly interested in influencing ontology development. Please send an e-mail to the first author if you are interested in joining this group, or visit https://groups.google.com/forum/#!forum/ahsontology.
- Regular webinars are planned to expose the developed content to members of the surveillance and scientific community, and gather objective feedback. Those interested can find information on http://datadrivensurveillance.org/ontology/.

Any interested members of the health surveillance community are encouraged to become engaged through interacting using any and all of these resources. As explained, no level of ontological knowledge will be needed to be an active member of the community, as the ontology curation group will oversee the task of translating the community input into iterations of ontology development, as well as translating the results back to the community.

6. Discussion and conclusions

We have discussed the development of an Animal Health Surveillance Ontological framework (AHSO), and presented a core structure modelling animal production. Active development is in place to expand this framework to model health observations from a varied number of contexts. Development is driven mostly from data examples, and content is added to the ontology by reusing other ontologies, reusing concepts from existing terminologies, and relying on surveillance and research community involvement to review and update knowledge.

Modelling the biological knowledge associated with animal health observations will support the automation of tasks related to translating those data into actionable information for surveillance (Awaysheh et al., 2017). In the case of syndromic surveillance, in particular, this should promote agreement regarding the definitions of syndromes; and allow these definitions to be used by machines across systems (regardless of the varied coding practices and languages used in different institutions), promoting interoperability and also knowledge sharing. Moreover, the model will facilitate simple and swift adjustment to accommodate new knowledge or the need to respond to new threats.

The AHSO framework is built under the assumption that we are not modelling actual health events, but rather modelling observations made about or relating to these events at specific moments in time. We may, for instance, need to identify observations related to the same animal or same herd, or observations connected to the same event. However, disease progression over time is not specifically modelled. The goal of extracting information relevant to the purpose of animal health surveillance is the main driver of the knowledge modelling process, though we will attempt to reflect all of the key interests of those contributing data and knowledge.

Chapman and collaborators (Chapman et al., 2010) highlighted the fact that while much of the biomedical knowledge necessary for surveillance is coded into ontologies within specific domains, such as ontologies of infectious diseases or anatomy, there existed no ontology to support syndromic classification for surveillance. The group then attempted to gather consensus definitions of syndromes within public health into an ontology, the syndromic surveillance ontology (SSO) (Okhmatovskaia et al., 2009), later extended into ESSO (Extended SSO) (Conway et al., 2011; Crubézy et al., 2005). Ultimately, the initiative did not move forward (personal communication with Wendy Chapman and Michael Conway) due to a lack of uptake within surveillance applications, and a lack of community involvement. In the AHSO framework, we have invested in a number of formats in an attempt to secure community engagement. We are planning yearly workshops to inform surveillance researchers, as well as regular webinars to engage community expertise. Those will be advertised in the project public pages listed above. Community involvement in its development and use has been pointed out as one of the key reasons for the success of the Gene Ontology (Bada et al., 2004).

Developers of other ontological frameworks have also emphasized the need to provide tools for ontology development, curation and adoption (Dhombres et al., 2017; Griffiths et al., 2017; Maurice et al., 2017), which are required to ensure ontology uptake. Once used in practice, the framework will facilitate the annotation of data using the ontological model (prospective); and enable mining of historical data empowered by the ontology (retrospective) (Chapman et al., 2011; Furrer et al., 2015). That is, information retrieval, integration and extraction will be empowered by complex semantic analysis based on the semantic relationships coded within the ontological model (Ferreira et al., 2012).

While highlighting the need to provide tools that can support the use of the ontology in practice, we should note that the adoption of semantic web tools does not require open access to the underlying data; rather they facilitate information extraction for those who have access to such data, and to promote interoperability (Ferreira et al., 2013). If tools are available which can query health data for syndromic classification, for instance, it is these tools that can be shared, rather than the data itself. Results of data analysis will, however, be comparable among systems from different institutions. Moreover, because the tools will be applicable to a larger range of datasets, without relying on the data being coded using the same standards, the tools for data analysis can be improved as a community effort. A community of syndromic surveillance researchers and practitioners can share knowledge and efforts to advance tools, which they then apply to their respective data sets privately.

The development of any ontology is a long-term task, but the growing number of biomedical ontologies and open access tools for ontology construction and management allow for the reuse of both knowledge and modelling solutions. The development of AHSO will build on achievements evident from the successful use of other ontologies. The methodology proposed is problem-oriented, collaborative, and will continue to promote community involvement.

Acknowledgements

This work is funded by Sweden's innovation agency (VINNOVA).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.prevetmed.2019.03.002.

References

Al Manir, M.S., Brenas, J.H., Baker, C.J.O., Shaban-Nejad, A., 2018. A surveillance infrastructure for malaria analytics: provisioning data access and preservation of interoperability. J. Med. Internet Res. 20. https://doi.org/10.2196/10218.

Allemang, D., Hendler, J., 2011. Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL. Morgan Kaufmann.

Awaysheh, A., Wilcke, J., Elvinger, F., Rees, L., Fan, W., Zimmerman, K., 2017. A review of medical terminology standards and structured reporting. J. Vet. Diagn. Invest., 104063871773827. https://doi.org/10.1177/1040638717738276.

Bada, M., Stevens, R., Goble, C., Gil, Y., Ashburner, M., Blake, J.A., Cherry, J.M., Harris, M., Lewis, S., 2004. A short study on the success of the gene ontology. Web Semant. Sci. Serv. Agents World Wide Web 1, 235–240. https://doi.org/10.1016/j.websem.2003.12.003.

Blomqvist, E., Hammar, K., Presutti, V., 2016. Engineering ontologies with patterns – the eXtreme design methodology. In: Hitzler, Pascal, Gangemi, Aldo, Janowicz, Krzysztof, Krisnadhi, Adila, Presutti, Valentina (Eds.), Ontology Engineering with Ontology Design Patterns. ISO Press.

Bodenreider, O., 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 32, D267–D270. https://doi.org/10.1093/nar/gkh061.

Buitelaar, P., Cimiano, P., Magnini, B., Brewster, C., 2005. Ontology Learning From Text: Methods, Evaluation and Applications. IOS Press.

Chapman, W.W., Dowling, J.N., Baer, A., Buckeridge, D.L., Cochrane, D., Conway, M.A., Elkin, P., Espino, J., Gunn, J.E., Hales, C.M., Hutwagner, L., Keller, M., Larson, C., Noe, R., Okhmatovskaia, A., Olson, K., Paladini, M., Scholer, M., Sniegoski, C., Thompson, D., Lober, B., 2010. Developing syndrome definitions based on consensus and current use. J. Am. Med. Inform. Assoc. 17, 595–601.

Chapman, W.W., Conway, M., Dowling, J.N., Tsui, F.-C., Li, Q., Chistensen, L.M., Harkema, H., Sriburadej, T., Espino, J.U., 2011. Challenges in adapting a natural language processing system for real-time surveillance. Emerg. Health Threats J. 4, 7–8 s68.

Cimino, J.J., 1999. From data to knowledge through concept-oriented terminologies: experience with the medical entities dictionary. J. Am. Med. Inform. Assoc. 7 (3), 288–297.

Conway, M., Dowling, J., Tsui, R., Chapman, W., 2011. Developing an application ontology for mining clinical reports: the extended syndromic surveillance ontology. Emerg. Health Threats J. 4, 15–16 s73.

Courtot, M., 2014. Semantic Models in Biomedicine: Building Interoperating Ontologies for Biomedical Data Representation and Processing in Pharmacovigilance. The University of British Columbia.

Crubézy, M., Connor, M.O., Buckeridge, D.L., Pincus, Z., Musen, M.A., 2005. Ontology-centered syndromic surveillance for bioterrorism a new trend : syndromic surveillance ontology-centered syndromic surveillance. IEEE Intell. Syst. 20, 26–35.

Dhombres, F., Maurice, P., Friszer, S., Guilbaud, L., Lelong, N., Khoshnood, B., Charlet, J., Perrot, N., Jauniaux, E., Jurkovic, D., Jouannic, J.-M., 2017. Developing a knowledge base to support the annotation of ultrasound images of ectopic pregnancy. J. Biomed. Semantics 8, 4. https://doi.org/10.1186/s13326-017-0117-1.

Dixon, B.E., Vreeman, D.J., Grannis, S.J., 2014. The long road to semantic interoperability in support of public health: experiences from two states. J. Biomed. Inform. 49, 3–8. https://doi.org/10.1016/j.jbi.2014.03.011.

Donald, M., Nzali, T., Aze, J., Bringay, S., Lavergne, C., Mollevi, C., Optiz, T., 2018. Reconciliation of patient/doctor vocabulary in a structured resource. Health Informatics J. 1. https://doi.org/10.1177/1460458217751014.

Dórea, F.C., Vial, F., 2016. Animal health syndromic surveillance: a systematic literature review of the progress in the last 5 years (2011–2016). Vet. Med. Reports 7, 157–169.

Dórea, F.C., Sanchez, J., Revie, C.W., 2011. Veterinary syndromic surveillance: current initiatives and potential for development. Prev. Vet. Med. 101. https://doi.org/10.1016/j.prevetmed.2011.05.004.

Dragisic, Z., Lambrix, P., Blomqvist, E., 2015. Integrating ontology debugging and matching into the eXtreme design methodology. Proceedings of the 6th Workshop on Ontology and Semantic Web Patterns.

Ferreira, J.D., Pesquita, C., Couto, F.M., Silva, M.J., 2012. Bringing epidemiology into the semantic web. CEUR Workshop Proc. 897, 1–5.

Ferreira, J.D., Paolotti, D., Couto, F.M., Silva, M.J., 2013. On the usefulness of ontologies in epidemiology research and practice. J. Epidemiol. Community Health 67, 385–388. https://doi.org/10.1136/jech-2012-201142.

Furrer, L., Küker, S., Berezowski, J., Posthaus, H., Vial, F., Rinaldi, F., Lenz, F., Küker, S., Berezowski, J., Posthaus, H., Vial, F., Rinaldi, F., Furrer, L., Küker, S., Berezowski, J., Posthaus, H., Vial, F., Rinaldi, F., Lenz, F., Küker, S., Berezowski, J., Posthaus, H., Vial, F., Rinaldi, F., 2015. Constructing a syndromic terminology resource for veterinary text mining. CEUR Workshop Proc. 1495, 61–70.

Gates, M.C., Holmstrom, L.K., Biggers, K.E., Beckham, T.R., 2015. Integrating novel data streams to support biosurveillance in commercial livestock production systems in developed countries: challenges and opportunities. Front. Public Health Serv. Syst. Res. 3, 74. https://doi.org/10.3389/fpubh.2015.00074.

Griffiths, E., Dooley, D., Graham, M., Van Domselaar, G., Brinkman, F.S.L., Hsiao, W.W.L., 2017. Context is everything: harmonization of critical food microbiology descriptors and metadata for improved food safety and surveillance. Front. Microbiol. 8, 1068. https://doi.org/10.3389/fmicb.2017.01068.

Lambrix, P., Ivanova, V., 2013. A unified approach for debugging is-a structure and mappings in networked taxonomies. J. Biomed. Semantics 4, 10. https://doi.org/10.1186/2041-1480-4-10.

Lambrix, P., Strömbäck, L., 2007. Where is my protein? Issues in information integration. Bioforum Eur. 7–8 24–26.

Lipscomb, C.E., 2000. Medical Subject Headings (MeSH). Bull. Med. Libr. Assoc. 88, 265–266. https://doi.org/10.4103/0019-5413.139827.

Maurice, P., Dhombres, F., Blondiaux, E., Friszer, S., Guilbaud, L., Lelong, N., Khoshnood, B., Charlet, J., Perrot, N., Jauniaux, E., Jurkovic, D., Jouannic, J.M., 2017. Towards ontology-based decision support systems for complex ultrasound diagnosis in obstetrics and gynecology. J. Gynecol. Obstet. Hum. Reprod. 46, 423–429. https://doi.org/10.1016/j.jogoh.2017.03.004.

Mirhaji, P., 2009. Public health surveillance meets translational informatics: a desiderata. J. Lab. Autom. 14, 157–170. https://doi.org/10.1016/j.jala.2009.02.007.

Mungall, C.J., Torniai, C., Gkoutos, G.V., Lewis, S.E., Haendel, M.A., 2012. Uberon, an integrative multi-species anatomy ontology. Genome Biol. 13, R5. https://doi.org/10.1186/gb-2012-13-1-r5.

Musen, M.A., 2005. Protégé: community is everything. Int. J. Hum. Stud. 62, 545–552. https://doi.org/10.1016/j.ijhcs.2005.03.002.

Noy, N.F., McGuinness, D.L., 2001. Ontology Development 101: A Guide to Creating Your First Ontology, Stanford Knowledge Systems Laboratory. https://doi.org/10.1016/j.artmed.2004.01.014.

Okhmatovskaia, A., Chapman, W., Collier, N., Espino, J., Buckeridge, D., 2009. SSO: the syndromic surveillance ontology. Proceeding Int. Soc. Dis. Surveillance.

Pesquita, C., Ferreira, J.J.D.J., Couto, F.M., Silva, M.M.J., 2014. The epidemiology ontology: an ontology for the semantic annotation of epidemiological resources. J. Biomed. Semantics 5, 4. https://doi.org/10.1186/2041-1480-5-4.

Poveda-Villalón, M., Gómez-Pérez, A., Suárez-Figueroa, M.C., 2014. OOPS! (OntOlogy Pitfall Scanner!). Int. J. Semant. Web Inf. Syst. 10, 7–34. https://doi.org/10.4018/ijswis.2014040102.

Presutti, V., Blomqvist, E., Daga, E., Gangemi, A., 2012. In: Suarez-Figueroa, M.C. (Ed.), Pattern-Based Ontology Design. Springer-Verlag, Berlin Heidelberg, pp. 35–64.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S., 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. 25, 1251–1255.

Stearns, M.Q., Price, C., Spackman, K.A., Wang, A.Y., 2001. SNOMED clinical terms: overview of the development process and project status. Proc. AMIA Symp. 662–666.

VanderWaal, K., Morrison, R.B., Neuhauser, C., Vilalta, C., Perez, A.M., 2016. Translating big data into smart data for veterinary epidemiology. Front. Vet. Sci. 4, 110. https://doi.org/10.3389/FVETS.2017.00110.

Xiang, Z., Courtot, M., Brinkman, R.R., Ruttenberg, A., He, Y., 2010. OntoFox: web-based support for ontology reuse. BMC Res. Notes 3, 1–12. https://doi.org/10.1186/1756-0500-3-175.